AP Stats – Class Notes – Section 3.2 – Day 3
The Role of s and r^2 in Regression

1. s is the **standard deviation of the residuals.** It estimates the size of a typical residual.

That is, s, measures the typical distance.....

S is sometimes called the.....

$$s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

2. r^2 is called the **Coefficient of Determination**. It measures the fraction of the variability in the $y - variable$ that is accounted for by the.....

More spefically:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

The coefficient of determination r^2 measures the percent reduction in the sum of squared residuals when using the least-squares regression line to make predictions, rather than the mean value of y.

In other words, r^2 measures the percent of the variability in the response variable that is accounted for by the least-squares regression line.

r^2 tells us how much better the LSRL does at predicting values of y

than simply....

Backpacking - Ninth-grade students at the Webb Schools go on a backpacking trip each fall. Students are divided into hiking groups of size 8 by selecting names from a hat. Before leaving, students and their backpacks are weighed. The data here are from one hiking group.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

Analyze the data using stapplet.com.

1. Using www.stapplet.com find the LSRL of the data. Write it below (in you know what form!)

2. Find and interpret ***S***, *the standard deviation of the residuals.*

3. Find and interpret the value of r^2, the coefficient of determination.

s and r^2

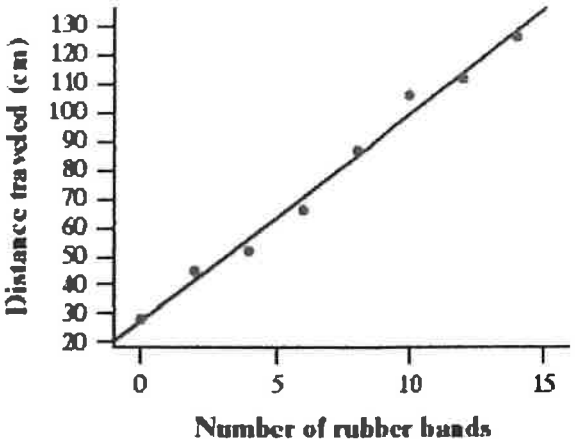
Big Ideas:

Mickey's last bungee jump *Interpreting s and r^2*

The class performed the "Mickey Mouse Bungee" activity. They connected rubber bands one at a time in a chain to Mickey's feet and then measured the distance that he traveled on his (last) bungee jump. The distance is measured from the edge of the jumping platform to the lowest point that Mickey's head reaches.

Here is the scatterplot of data from one of the groups with the regression line $\hat{y} = 27.42 + 7.21x$.

For this model, technology gives $s = 4.11$ and $r^2 = 0.989$.

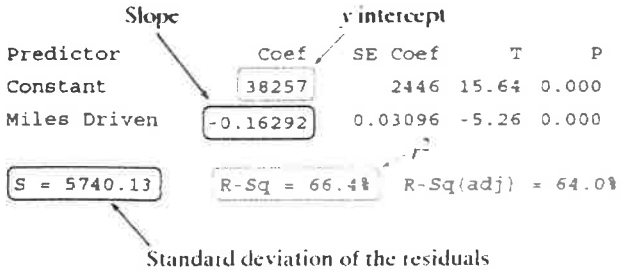


(a) Interpret the value of *s*.

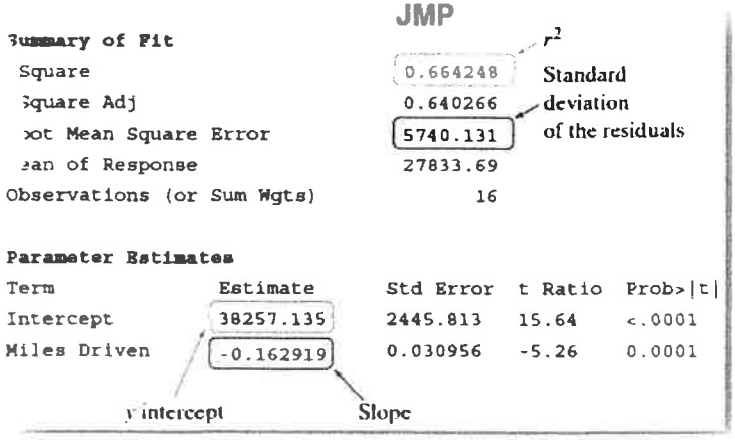
(b) Interpret the value of r^2.

Computer Output

Minitab



JMP

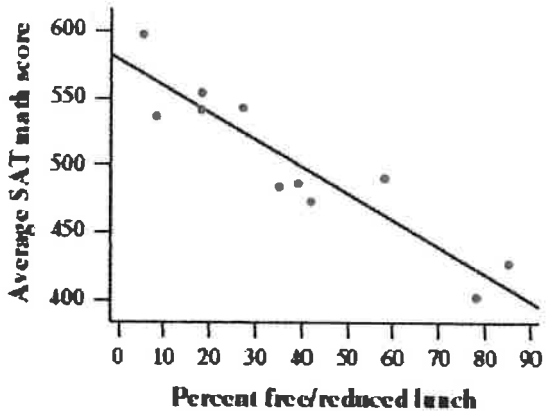


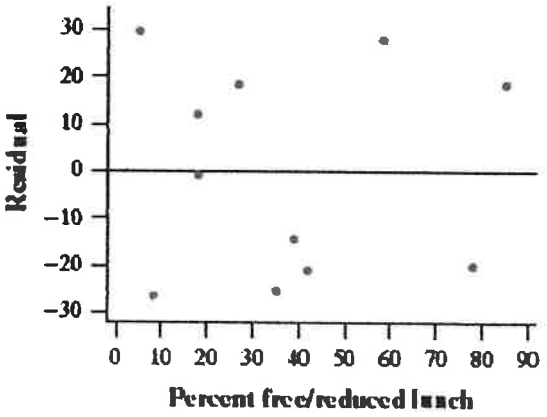
Can we predict a school's average SAT math score? *Interpreting regression output*

A random sample of 11 high schools was selected from all the high schools in Michigan. The percent of students who are eligible for free/reduced lunch and the average SAT math score of each high school in the sample were recorded.

Students with household income below a certain threshold are eligible for free/reduced lunch.

Here are a scatterplot with the least-squares regression line added, a residual plot, and some computer output:





Predictor	Coef	SE Coef	T	P
Constant	577.9	12.5	46.16	0.000
Foot length	−1.993	0.276	−7.22	0.000

S = 23.3168 R-Sq = 85.29% R-Sq(adj) = 83.66%

(a) Is a line an appropriate model to use for these data? Explain how you know the answer.

(b) Find the correlation coefficient.

(c) What is the equation of the least-squares regression line that describes the relationship between percent free/reduced lunch and average SAT math score? Define any variables that you use.

(d) By about how much do the actual average SAT math scores typically vary from the values predicted by the least-squares regression line with x = percent free/reduced lunch?